

This Page Is Inserted by IFW Operations  
and is not a part of the Official Record

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

## **IMAGES ARE BEST AVAILABLE COPY.**

As rescanning documents *will not* correct images,  
Please do not report the images to the  
Image Problem Mailbox.

(19)日本国特許庁(JP)

(12) 公開特許公報(A)

(11)特許出願公開番号

特開平5-265829

(43)公開日 平成5年(1993)10月15日

(51)Int.Cl. <sup>4</sup>	識別記号	庁内整理番号	FI	技術表示箇所
G 0 6 F 12/00	5 3 3 J	7232-5B		
11/20	3 1 0 E	7832-5B		
12/00	5 3 1 D	7232-5B		
	5 4 5 A	7232-5B		

審査請求 未請求 請求項の数8(全11頁)

(21)出願番号 特願平4-90248

(22)出願日 平成4年(1992)3月16日

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72)発明者 櫻庭 健年

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

(72)発明者 福澤 淳二

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

(72)発明者 黒田 澤希

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

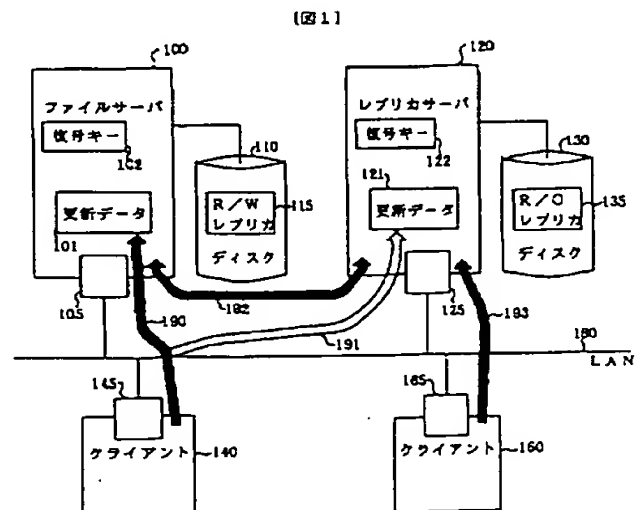
(74)代理人 弁理士 笹岡 茂 (外1名)

(54)【発明の名称】 レプリケートファイル更新方式

(57)【要約】

【目的】 分散データ処理システムにおける、レプリカサーバの読み出し専用レプリカを、新たに更新データの送信を行なうことなく効率よく最新化する。

【構成】 ファイルサーバ100にある、或るファイルの書き込み可能レプリカ115の更新を、通信ネットワーク180に接続されたクライアント140が通信路190を通じて行う時、該クライアントと前記ファイルの読み出し専用135を持つレプリカサーバ120との間の擬似的な通信路191を用いて同時にレプリカサーバで更新データを取り込み、該更新データを読み出し専用レプリカの最新化処理に用いることにより、当該最新化処理ではファイルサーバからレプリカサーバへの更新データ送信を不要とする。更にファイルサーバとレプリカサーバの間ではレプリカサーバが取り込んだ更新情報の妥当性を確認するための通信を行って、レプリカ115、と135の同一性を保証する。



## 【特許請求の範囲】

【請求項1】 通信ネットワークと、

当該通信ネットワークに接続された少なくとも1つのファイルサーバと少なくとも1つのレプリカサーバと、前記通信ネットワークに接続され前記ファイルサーバあるいはレプリカサーバの有するレプリケートファイルにアクセス可能な少なくとも1つのデータ処理システム

(クライアント) とからなる分散データ処理システムにおけるレプリケートファイル更新方式であって、

前記レプリカサーバは、前記通信ネットワークを介してレプリカサーバに対して行なわれる前記クライアントからのアクセスおよび該クライアントと前記ファイルサーバの間で行われる該ファイルサーバのレプリケートファイル更新のための前記通信ネットワーク上の交信を監視し、前記アクセスの検知および更新されるファイルサーバ内のレプリケートファイルと同一のファイルをレプリカサーバが有するか否かの検知をする検知手段と、

該検知手段により同一のファイルを有するという検知結果が得られた場合前記クライアントからファイルサーバに送信される更新内容を取り込む手段を備え、

前記レプリカサーバは、取り込んだ更新内容によりレプリケートファイルの更新を行うようにしたことを特徴とするレプリケートファイル更新方式。

【請求項2】 請求項1記載のレプリケートファイル更新方式において、

前記ファイルサーバとレプリカサーバは、取り込んだ更新内容によるファイル更新が何回目のファイル更新であるかを示すファイル更新回数を得る手段を備え、

前記レプリカサーバは、得られたファイル更新回数と前記ファイルサーバから送信されたファイル更新回数とを比較する更新回数比較手段を備え、

前記レプリカサーバは、前記更新回数比較手段による比較の結果、ファイル更新回数が一致したとき取り込んだ更新内容によりレプリケートファイルの更新を行うようにしたことを特徴とするレプリケートファイル更新方式。

【請求項3】 請求項2記載のレプリケートファイル更新方式において、

前記レプリカサーバは、前記更新回数比較手段による比較の結果、ファイル更新回数が一致しないとき、更新内容送信要求を前記ファイルサーバに送信し、該ファイルサーバから送信された更新内容を受信し、受信した更新内容によりレプリケートファイルの更新を行うようにしたことを特徴とするレプリケートファイル更新方式。

【請求項4】 請求項2記載のレプリケートファイル更新方式において、

前記ファイルサーバとレプリカサーバは、取り込んだ更新内容の妥当性を示す妥当性判定用データを生成する生成手段と、

前記レプリカサーバは、得られた妥当性判定用データと

前記ファイルサーバから送信された妥当性判定用データとを比較する妥当性比較手段を備え、

前記レプリカサーバは、前記妥当性比較手段による比較の結果、妥当性判定用データが一致したとき取り込んだ更新内容によりレプリケートファイルの更新を行うようにしたことを特徴とするレプリケートファイル更新方式。

【請求項5】 請求項4記載のレプリケートファイル更新方式において、

10 前記レプリカサーバは、前記更新回数比較手段による比較の結果と妥当性比較手段による比較の結果の両方またはいずれか一方が不一致のとき、更新内容送信要求を前記ファイルサーバに送信し、該ファイルサーバから送信された更新内容を受信し、受信した更新内容によりレプリケートファイルの更新を行うようにしたことを特徴とするレプリケートファイル更新方式。

【請求項6】 請求項1乃至請求項5のいずれかの請求項記載のレプリケートファイル更新方式において、

20 前記検知手段を、前記アクセスの検知をする第1の検知手段と前記更新されるファイルサーバ内のレプリケートファイルと同一のファイルをレプリカサーバが有するか否かの検知をする第2の検知手段に分割して独立に設けたことを特徴とするレプリケートファイル更新方式。

【請求項7】 請求項6記載のレプリケートファイル更新方式において、

30 前記レプリカサーバに少なくとも2つのプロセッサを設け、前記クライアントからのアクセスに対する処理と、前記更新されるファイルサーバ内のレプリケートファイルと同一のファイルに対するレプリカサーバにおける処理を夫々別々のプロセッサにより行なうようにしたことを特徴とするレプリケートファイル更新方式。

【請求項8】 請求項1乃至請求項7のいずれかの請求項記載のレプリケートファイル更新方式において、

40 前記レプリカサーバは、前記クライアントから前記ファイルサーバに送信される暗合化された情報を解読するのに必要な暗合複合情報を格納する格納手段を設け、予め前記ファイルサーバから前記レプリカサーバに送信される暗合複合情報を該格納手段に格納し、前記クライアントから前記ファイルサーバに送信される更新内容を取り込むとき前記暗合複合情報を用いて取り込むようにしたことを特徴とするレプリケートファイル更新方式。

【発明の詳細な説明】

【0001】

【産業上の利用分野】 本発明は分散データ処理システムにおけるレプリケートファイルの効率の良い運用に好適なレプリケートファイル更新方式に関する。

【0002】

【従来の技術】

(1) スヌープキャッシュ

50 スヌープキャッシュについては、例えば、丸善株式会

社、MAT、電子情報通信編、並列処理機構（1989年）、PP167-207に記載されている。

【0003】各プロセッサが別々のキャッシュメモリを備え、かつ共通の主記憶にアクセスするように構成されたマルチプロセッサシステムでは、主記憶上の同一のアドレスにあるデータを各プロセッサのキャッシュメモリにコピーしてアクセスする。この場合、あるプロセッサが主記憶を更新すると、更新したプロセッサ以外のプロセッサが、そのキャッシュメモリに更新されたアドレスのデータを持っているときは、そのデータを無効化するなどして、キャッシュ間の無矛盾性（コヒーレンシー）を保証する必要がある。スヌープキャッシュはそのためのキャッシュ制御方式、ないしその制御方式を採用するキャッシュメモリのことである。

【0004】スヌープキャッシュでは各プロセッサがバス結合されているマルチプロセッサシステムにおいて、1つのプロセッサが主記憶を更新すると、各プロセッサのキャッシュ制御回路はバスを監視することによって更新アドレスを知ることができ、該当するデータをキャッシュメモリに持っていればそれを無効化してコヒーレンシーを保つ。スヌープキャッシュでは各プロセッサが監視するのは、それと同等の機能を持つ他のプロセッサのメモリへの書き込みである。上記文献PP196-197では、書き込み放送型のキャッシュ更新プロトコルについて解説している。このプロトコルでは他のプロセッサがメモリ更新するデータが自らのキャッシュ内にある場合、そのキャッシュを無効化するのではなく、データを該当するキャッシュに直ちに書き込む。スヌープキャッシュ適用システムではバスの信頼性が高く、1プロセッサによるメモリ更新は他のプロセッサに確実に伝えられ、各スヌープキャッシュにおけるコヒーレンシーの保証動作が確実に起動される。

#### 【0005】(2) 分散ファイルシステム

分散データ処理システムは、LAN (Local Area Network) と呼ばれる通信ネットワークにパーソナルコンピュータやワークステーションなどのデータ処理システムを複数接続し、相互に通信ネットワークを介して情報の授受を行いながら、それぞれの処理を遂行するシステムである。分散データ処理システムにおける重要な技術として、通信ネットワーク上の1つのデータ処理システムから他のデータ処理システムに存在するファイルのアクセスを可能とする分散ファイルシステムがある。分散ファイルシステムは、1つのデータ処理システムに存在するファイルを複数のデータ処理システムが共用することを可能とする。ここでは、共用されるファイルが存在するデータ処理システムをファイルサーバと呼び、ファイルサーバ上の共用ファイルにアクセスするデータ処理システムをクライアントと呼ぶことにする。

【0006】特定の共用ファイルが頻繁にアクセスされ

るとそのファイルサーバに要求が集中し、分散データ処理システム全体の性能上のボトルネックとなる可能性がある。そこでこのファイルのコピーを第2のファイルサーバにも置き、このファイルへのアクセス要求の一部を第2のファイルサーバで処理することにより、2つのファイルサーバの間で負荷分散し、前記のボトルネックの解消を図ることが行われる。同一ファイルのコピーを複数のデータ処理システムに置くことにより、ファイル破壊が発生してもコピーを代替として用いることにより、当該ファイルへのアクセスが中断せず、また回復も容易となるので、分散データ処理システム全体の信頼性向上にもなる。同一ファイルのコピーの各々をそのファイルのレプリカ、あるいはレプリケートファイルと呼び、コピーを作成することレプリケーションという。

【0007】分散ファイルシステムについてはエー・シー・エム、コンピューティングサーベイズ、第22巻、第4号（1990年）PP321-374 (ACM, Computing Surveys, Vol. 22, No. 4 (1990) PP321-374) において解説されており、レプリケーションについては同文献のPP339-340に論じられている。上記文献にもある通り、レプリケーションでは複数あるレプリカの更新方式が問題である。更新が発生した時点で全てのレプリカに同様の更新を行う方式はレプリカ間の同等性の保証はできるがオーバーヘッドが大きい。一方、レプリカ間の同等性を全く保証しない方式は可用性、及び性能の観点からは良好であるが、同等性がくずれるため、処理結果の正当性が保証されない。よく行われる方法は書き込み可能なレプリカ1つと読み出し専用のレプリカを複数持ち、書き込み可能レプリカに対する更新内容を読み出し専用レプリカのあるデータ処理システムに後から転送して、読み出し専用レプリカを最新化する、即ち書き込み可能レプリカの最新状態を読み出し専用レプリカにアトミック性を保証しながら反映させる方法である。一時的にレプリカ間の同等性がくずれるが、適当な時間経過の後は同等性が回復され、ファイル使用中はファイル内部の一貫性が保証されるので使用に耐え得る方式となっている。特に読み出し専用ファイル、ないし更新が極めてまれなファイルに対して好適である。

【0008】以下では記述の簡単化のために書き込み可能レプリカが存在するファイルサーバを単に（そのファイルの）ファイルサーバと呼び、読み出し専用レプリカが存在するファイルサーバを（そのファイルの）レプリカサーバと呼ぶことにする。ファイルサーバ、及びレプリカサーバという呼び方は特定のファイルを決めたときに定まる概念であり、物理的に同じデータ処理システムが特定ファイルのファイルサーバであると同時に、別のファイルのレプリカサーバであることは可能である。

#### 【0009】

【発明が解決しようとする課題】書き込みが頻繁なファ

イルの場合は、読み出し専用レプリカの最新化处理が頻発し、最新化处理に伴う更新データ転送量が多くなる。そのため、書き込み可能レプリカを持つファイルサーバでは更新ファイル、ないし更新データの送信処理のオーバーヘッドが発生し、また通信路に新たな負荷が発生してそのビジー率が上がる。1つのファイルに対し、複数のレプリカを持った場合には、持たない場合に比べ、これらのオーバーヘッドはレプリカの個数倍だけ増大する。

【0010】ファイルサーバのオーバーヘッドは複数のファイルサーバを導入して負荷分散することにより緩和できる。しかし、通信路は分散データ処理システムに唯一の資源であり、その負荷の緩和は容易ではない。分散データ処理システムを評価する1つの尺度に規模拡張性(スケーラビリティ)があるが、通信路はそのボトルネックの1つである。即ち、レプリケーションによるファイルの可用性、及び信頼性の向上とシステムの規模拡張性の間にはトレードオフが存在し、その解決方式が問題である。本発明の目的は、ファイルサーバにおけるレプリカ更新のための更新データ送信オーバーヘッド、及び通信路の負荷を発生させることのないレプリケートファイルの効率の良い、かつ高信頼な更新方式を提供することにある。

#### 【0011】

【課題を解決するための手段】上記目的を達成するために、本発明では書き込み可能レプリカの更新に伴う、ファイルサーバとクライアントとの通信を、読み出し専用レプリカを持つデータ処理システムであるレプリカサーバから監視し、レプリカサーバは書き込みを認知したら書き込み内容を取り込んで読み出し専用レプリカの最新化处理を行う。また、レプリカサーバの読み出し専用レプリカの更新に際して、更新内容の妥当性を更新対象ファイルの更新回数あるいは妥当性判定用データにより判定し、この判定を、レプリカサーバで得られた更新対象ファイルの更新回数あるいは妥当性判定用データとファイルサーバで得られた更新対象ファイルの更新回数あるいは妥当性判定用データとの一致をファイルサーバとの通信により確認することにより行なう。レプリカサーバは、妥当性が確認された場合は、読み出し専用レプリカの最新化、即ち更新を行い、エラーが検出された場合は、ファイルサーバから改めて最新状態のレプリカを送信させ、これにより読み出し専用レプリカを更新する。

#### 【0012】

【作用】レプリカサーバは、読み出し専用レプリカの最新化のための更新データとして、ファイルサーバにおける書き込み可能レプリカの更新処理時に授受された更新情報を取り込み、これを用いて読み出し専用レプリカを更新する。このため、読み出し専用レプリカの最新化のために新たにファイルサーバからレプリカサーバに更新情報を送る必要がなく、従って通信ネットワークの負荷を抑えることができる。また、読み出し専用レプリカの

最新化に必要な更新情報の特徴をよく表し、かつ更新情報全体よりかはるかに短いデータである更新対象ファイルの更新回数あるいは妥当性判定用データをファイルサーバから受け取って、あるいはファイルサーバに送って、両サーバで得られた更新回数あるいは妥当性判定用データの一致を調べ、データの妥当性を確認する。これにより、通信ネットワーク等で生じるデータ誤りに対処できる。妥当性が確認された場合は上記同様、新たな更新情報の送信を行わずに読み出し専用レプリカの最新化を実行する。確認に用いるデータは極めて小さいので通信ネットワークの負荷の増加は問題にならない。妥当性確認の結果、更新データの不当性(エラー)が認められた場合は、書き込み可能レプリカの全内容、ないし最新化に必要な更新情報をファイルサーバからレプリカサーバに送信して書き込み専用レプリカの最新化を実行する。この場合はファイルサーバ、及び通信ネットワークにおける負荷の緩和は実現できないが、エラーの発生率は極めて小さいと考えられるため、十分な負荷削減効果があり、かつ本発明を適用しない場合と同程度の信頼性を確保することができる。

#### 【0013】

【実施例】以下、図を用いて本発明の実施例を説明する。図1は本発明が適用可能な分散データ処理システムの概要を示した構成図である。書き込み可能なレプリケートファイル(レプリカ)を備えるデータ処理システムであるファイルサーバ100、読み出し専用のレプリケートファイルを備えるデータ処理システムであるレプリカサーバ120、第1のクライアント(ファイルサーバあるいはレプリカサーバの有するレプリケートファイルにアクセス可能なデータ処理システム)140、第2のクライアント160はそれぞれ独立したデータ処理システムであり、各データ処理システムは共通の通信ネットワーク180にそれぞれインターフェースボード105、125、145、及び165により接続されており、相互に通信、及びデータ転送が可能である。例えばファイルサーバ100、及びレプリカサーバ120は高性能なワークステーション(WS)、第1、第2のクライアントは性能は並みで安価なWS、あるいはパーソナルコンピュータ(PC)からなり、通信ネットワークとしてローカルエリアネットワーク(LAN)を使用するのが典型的な構成の例である。ファイルサーバ100、及びレプリカサーバ120はそれぞれファイル格納のためにディスク装置110、及び130を有しており、そこに格納されたファイルは、クライアント140、160から共通にアクセスすることができる。特定のファイルを固定して考えるとき、そのファイルの書き込み可能(R/W)レプリカ115はファイルサーバ100のディスク110に格納され、そのファイルの読み出し専用(リードオンリ:R/O)レプリカ135はレプリカサーバ120のディスク130に格納されている。クライ

アント140はこのファイルをファイルサーバ100上のレプリカ115を読み出し、クライアント160は同じファイルをレプリカサーバ120上のレプリカ135を読み出すことにより、ファイルサーバとレプリカサーバにファイル転送処理オーバーヘッドを分散させることができる。

【0014】レプリカ115は書き込み可能レプリカであり、クライアント140はこれの更新を要求することができる。更新情報101、即ちファイルの更新後の新しいデータはクライアント140から物理的な通信路145、180、及び105から構成される通信路190を経て伝達される。ファイルサーバ100では更新要求に従い、このファイルの書き込み可能レプリカ115を更新する。この時点でレプリカ115とレプリカ135の間には内容的な不一致が生じており、将来、適当な時点に何らかの方法で両者を一致させる必要がある。

【0015】通信ネットワークを介した通信は送信側、及び受信側の交信手順を定めたプロトコルに乗っ取って行われ、この中には受信すべきデータ処理装置の指定方法、即ちアドレスの指定方法も含んでいる。データ転送自体は通信路上に電気的な振動を発生させ、あるいはそれを検出することにより行われ、通信は、各データ処理システムが検出した電気振動に含まれるアドレス情報を読み取って、自らへの通信であるか否かを判定することにより成立する。

【0016】本実施例ではレプリカサーバ120のインターフェースボード125においてレプリカサーバ120への本来の通信だけでなく、ファイルサーバ100への通信も取り込み、書き込み可能レプリカの更新とその更新内容を知る。これは通信ネットワーク180上の電気振動はインターフェースボード105と同様、インターフェースボード125においても検出されるので可能である。即ち、クライアント140からレプリカサーバに到る通信路191が事実上あって、これを通じてファイルサーバへの更新データの送信を監視する事を可能とし、得られた更新情報121を読み出し専用レプリカ135の最新化に使用する。

【0017】分散データ処理システムでは通信のセキュリティのために通信内容を暗号化することがある。ファイルサーバ100はその複合キー102をあらかじめ知っているため交信データ101の交信を行う事ができるが、レプリカサーバではデータを取り込んでも、そのままではそれを正しく更新情報として使用することはできない。そこでファイルサーバ100はレプリカサーバ120にあらかじめ複合キー102を送信しておく必要がある。レプリカサーバ120では当該複合キーを自らのシステム内に複合キー122として保持する。取り込んだ更新データ121の複合にはこの複合キー122を用いて正しくレプリカの更新を行うことができる。

【0018】図2はレプリカサーバ120におけるファ

イルサーバの通信を取り込むためのインターフェースボード230をレプリカサーバ120への通信を受け付ける本来のインターフェースボード210の他に設けて、より確実な通信監視を行うようにした他の実施例を示している。

【0019】通信ネットワーク180とのインタフェースであるインターフェースボード210は電氣的接続の他に、通信のあて先アドレスの処理を行う。即ち、指定アドレスが自らのアドレスであった場合は、通信データを取り込み、その処理をデータ処理システム120に行わせるためにデータ処理システム120のプロセッサ260に割り込み280をかける。通信データの処理中にファイルサーバで更新が行われると、レプリカサーバ120のプロセッサ260はその更新データの取り込みに失敗する可能性がある。そこで、ファイルサーバ100の通信を監視するために、レプリカサーバ120本来のインターフェースボード210とは別に、ファイルサーバ100の通信監視専用のインターフェースボード230を設ける。さらにインターフェースボード230が取り込んだデータを確実に処理するため、専用のプロセッサ250を設け、監視データの処理の場合は常にプロセッサ250に割り込むように構成すると、より良好な性能を実現できる。クライアント140がファイルサーバ100と交信すると事実上、物理的な通信路145、180、及び230からなる通信路291が構成される。これは図1の191の通信路に相当するものである。

【0020】プロセッサ250とプロセッサ260の関係は以下のものである。すなわち、プロセッサ250がファイルサーバにおけるファイルの書き込み可能レプリカ更新を監視し、その更新内容を取り込んでいる間はプロセッサ260はレプリカを用いたサービスを含む一般の処理を実行し続けることができる。読み出し専用レプリカのバージョンアップ処理では、分散ファイルシステムの制御方針にもよるが、新旧レプリカの切り替え処理とファイルサーバにおけるファイルの更新が並行して実行される場合がある。この更新はレプリカサーバにおいて作成中の新レプリカの次のバージョンに反映されるべきものであり、更新内容の取り込みを引続き行なう必要がある。更新内容取り込み専用プロセッサを用いた複数のプロセッサによって構成されるレプリカサーバの場合、レプリカのバージョンアップ処理は一般処理用プロセッサ260が新レプリカを作成し、プロセッサ250はレプリカ更新監視を続行する。そのため、バージョンアップ処理が起動される段階で、更新データバッファの切り替えと同時にそれまでの更新データバッファのプロセッサ250からプロセッサ260への引継ぎが行なわれ、それに伴う、両プロセッサ間での同期処理が必要となる。

【0021】次にレプリカサーバ120でのファイル交信データの処理について説明する。レプリカサーバ12

0のインターフェースボード210、230は通信ネットワーク180で通信が行われると通信データを取り込むと同時にレプリカサーバ120のプロセッサに割込みをかけてその処理を要求する。図3はこのときの割込み処理を示している。割込み処理300では、ステップ310にて前記取り込まれたデータの内容、ないし割込み情報から、当該データが、当レプリカサーバに読み出し専用レプリカがあるファイルに係わるものでありかつ書き込み可能レプリカを持つかのファイルサーバとの通信に係わるものであるかを調べる。該当するファイルサーバの通信である場合は、ステップ320に進む。

【0022】ステップ320ではその通信内容がファイル更新に係わるものかを調べ、ファイル更新用の通信ならば、ステップ330に進む。ステップ330では更新対象であるファイルが、当レプリカサーバに読み出し専用レプリカがあるファイルであるかを調べる。もし、レプリカが存在するファイルならばステップ350に進む。ステップ350では前記データをレプリカ更新データとして取り込み、保存する。更にステップ360にて、レプリカ、及びレプリカ更新データの管理情報を作成、ないし更新し、後の読み出し専用レプリカの最新化処理で用いる。

【0023】次にステップ310、320、及び330の判定処理で用いるデータについて説明する。図4は割込み時にメモリ上に取り込まれている通信データの構造の例を示している。当該データはそれを参照する処理のレベルに対応して入れ子構造になっている。データ構造400は通信ネットワーク上を伝送される時のデータ構造である。発信元アドレス401、あて先アドレス402、データタイプ403、エラー検出用冗長データ405、及びデータ本体を含んでいる。

【0024】ステップ310では、あて先アドレス402を参照することにより、当該通信データが関係するファイルサーバへの通信であるかを判定できる。データ構造400のデータ本体404は上位レベルの処理で用いるデータ構造420を含んでいる。データ構造420は発信元ID421、あて先ID422、制御情報423、及び上位処理に渡すデータ424を含む。あて先ID422はその通信を受け取るべき機能、ないしプログラムを表す。レプリカサーバ120はファイルサーバ100の制御プログラムがどのようなあて先IDを持つかを知らしており、このフィールド422により、ファイルサーバに対する通信であることが分かる。

【0025】データ424は440に示すデータ構造を持つ。サービスID441にはファイルサーバの制御プログラムに対するサービス要求内容が表示される。当該フィールドを参照することにより、ファイルの更新処理か否かが分かる。従って、ステップ320ではフィールド422、441を参照してファイル更新要求であるこ

とを判定する。ファイル更新要求である場合はデータ構造440の後続のフィールドには、更新対象となるファイルのファイルID442、更新位置の相対アドレス443、更新範囲を表す更新長444、及び更新内容そのものである更新データなどがある。レプリカサーバでは格納しているレプリカの管理テーブルを所有しており、ステップ330ではファイルID442が示すファイルのレプリカを持っているか否かを判定することができる。更新内容を表すフィールド443、444、445はステップ350で用いられ、内容の一部はディスクに格納される。

【0026】次に更新データの妥当性について説明する。レプリカの更新データはレプリカサーバ120とファイルサーバ100で別々に取り込まれ、更新に使われる。特にファイルサーバ100では更新要求をしたクライアント140との間で適当なプロトコルに乗っ取り、確実なデータ転送が行われるが、レプリカサーバではこの通信を監視するのみなので、取り入れた更新データの妥当性は必ずしも保証できない。データを取り入れることができた場合には、データにはエラー検出用の冗長データが付加されており、データの妥当性はある程度保証できる。しかし、例えばレプリカサーバが過負荷のため、ファイルサーバ100の通信の監視にもれが生じ、更新データが失われる可能性がある。従って、何らかの手段により、レプリカサーバが取り込んだ更新データの妥当性チェックを、ファイルサーバ100を交えて行う方が望ましい。

【0027】各取り込みデータのエラー検出用冗長データの有効性を仮定すれば、毎回の当該データのチェックの他に、ファイルサーバ、及びレプリカサーバにおける更新回数の一致を確認すればレプリカサーバにおいて取り込んだ更新データの妥当性は保証されたと考えることができる。すなわち、例えばファイルサーバ及びレプリカサーバに格納されているファイルAの或る時点以降の更新について見ると、この時点以降の最初の更新を1回目の更新とし、以下更新が或る度に2回目、3回目、…、n回目の更新とする。各回の更新はファイルA全体の更新であってもよく、また、1回目の更新ではファイルAの前半部分の20バイト、2回目の更新ではファイルAの後半部分の30バイト、3回目の更新ではファイルAの中央部分の40バイトというようなファイルAの部分的更新であってもよい。

【0028】一方、エラー検出用冗長データの有効性が不十分であったり、回数的一致だけでは妥当性に不安がある場合には、更新データに依存した妥当性判定用のデータをレプリカサーバ、及びファイルサーバの双方で作成し、両者の一致を確認する必要がある。妥当性判定のための通信は更新されたデータよりはるかに小さいことが、本発明の目的である通信ネットワークの負荷を抑えるための条件である。妥当性判定用データとして、各回

の更新データのエラー検出用冗長データを送り、両者を突き合わせる。これにより、不一致を検出するだけでなく、不一致の場合には、エラー検出用冗長データの不一致が発生した回の更新データのみをレプリカサーバに送るようにすれば、全ての更新データを送る場合に比べて、データ転送量を削減することができる。

【0029】図5は、この妥当性判定用データの構造を示している。妥当性判定用データ500は、前回のバージョンアップ以降から数えて1回目のから最新の更新までの更新データに付随するエラー検出用冗長データフィールド510、520、……、530からなる。例えば2回目の更新時の更新データのみをレプリカサーバに転送するのみでレプリカの同一性を保証することができる。また、別の妥当性判定の仕方として、ファイルサーバ及びレプリカサーバの双方でn回目の更新が行われた後の更新内容についてのエラー検出用冗長データを生成し、エラー検出用冗長データのみをファイルサーバからレプリカサーバに判定用データとして転送し、その一致を確かめるようにしてもよい。これによれば、途中でエラーが発生しても、また更新データの取り込みが欠落しても、ファイルサーバからn回目までの更新データが送られ書き換えが行なわれることにより、データの上書きが行なわれ、最終的に内容が一致すればレプリカとして有効であるため、エラーのペナルティを軽減する効果がある。特に、この場合、更新回数を更新データの妥当性判定のためのデータとして用いなければならないという条件を外すことができる。

【0030】次にファイルサーバで用いられるレプリカ管理テーブルについて説明する。図6はファイル管理エントリ600とそのレプリカを管理するレプリカ管理エントリ620の構造を示している。管理対象となるファイル、及びレプリカは複数存在するのでいずれもチェーン構造を伴っている。ファイル管理エントリ600(FME)には当該エントリが管理するファイルのファイルID601、当該ファイルのレプリカのレプリカ管理エントリ(RME)へのポインタ603を含む。レプリカ管理エントリ620にはレプリカID621、当該レプリカが存在するデータ処理装置のアドレス623、当該レプリカの現在の状態を表すカレントバージョン624、当該バージョンをレプリケートしてから後の当該ファイルの更新回数を示す更新カウント625、更に更新経過を特徴づける妥当性判定用データ626を含む。

【0031】次にレプリカサーバにおける読み出し専用レプリカの最新化処理について説明する。図7はレプリカサーバにおける読み出し専用レプリカの最新化処理の1実施例におけるファイルサーバにおける処理700、及びレプリカサーバにおける処理750を示すフローチャートである。本図には両サーバ間の通信を表すかぎ付き矢印711、731、771、及び772が示されている。また通信711で送信されるデータ720も示さ

れている。本実施例では、読み出し専用レプリカの最新化処理はファイルサーバから起動するものとする。

【0032】まず、ファイルサーバがステップ710にてレプリカサーバにその最新化要求を発行する(711)。この通信711では送信データ720として、最新化すべきレプリカのファイルID721、前回最新化してから後の、当該ファイルの書き込み可能レプリカの更新回数722、更に必要ならば更新データの妥当性判定用データ723などが送られる。更新回数722、及び判定用データ723はファイルサーバにおけるレプリカ管理エントリ620内の更新カウント625、及び判定用データ626の値が用いられる。本通信711により、レプリカサーバにおけるレプリカ最新化処理750が起動される。レプリカサーバではステップ755において送信された更新回数722と判定用データ723とレプリカサーバ内部で計数、あるいは生成した更新回数、及び判定用データとを比較して、レプリカサーバにおける更新データの妥当性を評価する。妥当性が確認されたならば、ステップ760に進む。

【0033】ステップ760ではレプリカサーバが取り込んだ更新データを用いて最新化処理を当該レプリカサーバ内の読み出し専用レプリカに対して行う。最新化処理が成功した場合はステップ762でファイルサーバからの最新データの送信が不要であることを送信771し、ステップ780にてレプリカ管理情報を更新してレプリカ最新化処理を終了する。また最新化処理が失敗した場合はステップ770に進む。ステップ755におけるレプリカサーバが取り込んだ更新データにの妥当性が確認されなかった場合、及びステップ761で最新化処理が失敗したと判断された場合は、ステップ770にて、ファイルサーバに対して指定ファイルの最新版の送信を要求する応答772を返す。

【0034】ファイルサーバはステップ715でレプリカサーバからの応答771、772を調べ、最新バージョンの送信要求でない、即ちレプリカサーバにおけるレプリカの最新化が完了した場合はステップ740でレプリカ管理情報の更新を行い、処理を終了する。また、ステップ715で最新バージョンの送信が要求されていることが分かった場合には、ステップ730に進み、最新バージョンの送信731を行う。処理は再びステップ720に戻り、レプリカサーバにおけるレプリカの最新化が完了したことを確認し、ステップ740を経て終了する。レプリカサーバでは最新バージョンの転送731により入手した更新データを用いて、ステップ775でレプリカの最新化処理を行い、ステップ761へ進む。ステップ762にて最新化処理終了を応答771して終了する。

【0035】次に、本発明とスヌープキャッシュにおけるキャッシュコヒーレンシーの保証方式との違いについて説明する。スヌープキャッシュでは、データ更新元は



プロセッサであり、バスを監視するのは更新元と同様の機能を持つ他のプロセッサであり、更新対象はメモリであって、前記プロセッサとは異なるレベルの実体である。一方、本発明ではファイルの更新元はクライアントであり、通信路を監視するのはレプリカサーバであり、更新されるのはレプリカサーバと同列の機能を持つファイルサーバにある書き込み可能レプリカである。即ち、監視を行うのが更新要求元と同等か、更新対象と同等かという構成上の根本的な違いがある。また、スヌープキャッシュでは、キャッシュ更新は監視の結果取り込まれた更新データを直ちにキャッシュに反映する。一方、本発明では監視の結果取り込んだ更新データは適当な記憶手段に蓄えられ、データ更新が複数回行われた後でも読み出し専用レプリカの最新化を行うことが可能である。即ち、取り込んだデータの反映の方法に大な違いがある。更に、本発明では、データの最新化にあたり、取り込んだ更新データの妥当性をファイルサーバとの通信を通じて確認することにより、信頼性の高い、レプリケーションを実現している。

#### 【0036】

【発明の効果】本発明によれば、書き込み可能レプリカの更新と同時にレプリカサーバにもその更新内容が取り込まれるので、読み出し専用レプリカを最新化するにあたり、改めてファイルサーバからレプリカサーバに更新内容を送信する必要がなく、ファイルサーバ、及び通信ネットワークの負荷を増大させることのないレプリケートファイルの最新化が可能となる。更に本発明では、レプリカの更新の際に更新情報の妥当性をファイルサーバとの交信を交えて確認するのでレプリカ間の同一性を高いレベルで保証することが可能であり、信頼性の高いレプリケーションが実現できる。

#### 【図面の簡単な説明】

【図1】本発明の1実施例における適用対象である分散データ処理システムの概要を示す構成図である。

【図2】図1に示されたレプリカサーバの他の実施例の構成を示す構成図である。

【図3】本発明の1実施例におけるレプリカサーバでのデータ取り込み処理を示すフローチャートである。

【図4】本発明の1実施例におけるレプリカ取り込み処理に用いられるデータを示す説明図である。

【図5】本発明の1実施例における妥当性判定用データの構造を示す図である。

【図6】本発明の1実施例におけるファイル、及びレプリカ管理データを示す説明図である。

【図7】本発明の1実施例におけるレプリカ最新化処理を示すフローチャートである。

#### 【符号の説明】

100 ファイルサーバ（データ処理システム）  
101 更新データ  
102 暗号復号キー

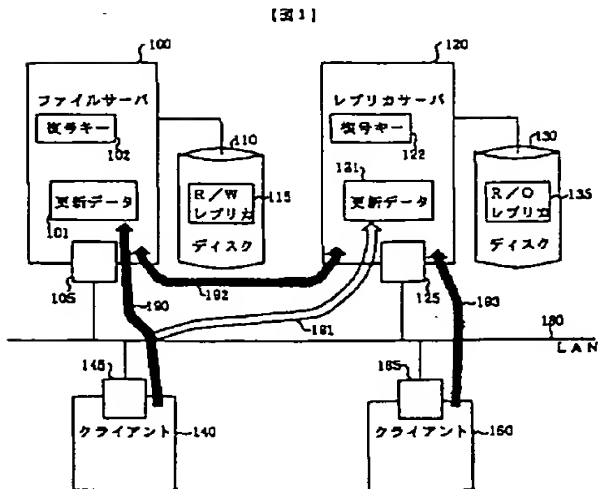
105、125、145、165 インターフェースボード  
110 ディスク  
115 書き込み可能レプリカ  
120 レプリカサーバ（データ処理システム）  
121 更新データ  
122 暗号復号キー  
130 ディスク  
135 読み出し専用レプリカ  
140、160 クライアント（データ処理システム）  
180 通信ネットワーク（LAN）  
190 クライアントとファイルサーバ間の通信  
191 クライアントとレプリカサーバ間の擬似的な通信  
192 ファイルサーバとレプリカサーバ間の通信  
200 ファイルサーバと通信ネットワークを接続するインターフェースボード  
210 レプリカサーバと通信ネットワークを接続する第1のインターフェースボード  
230 レプリカサーバと通信ネットワークを接続する第2のインターフェースボード  
250 第2のインターフェースボードからの割込みを処理するプロセッサ  
260 第1のインターフェースボードからの割込みを処理するプロセッサ  
270及び280 割込み信号  
290 クライアントとファイルサーバ間の通信  
291 第2のインターフェースボードクライアントとレプリカサーバ間の擬似的な通信  
400、420、440 通信情報及びその詳細情報  
402 宛先アドレス  
422 宛先識別子  
441 要求サービス指定子  
442 ファイル識別子  
443 更新箇所の相対アドレス  
444 更新データの長さ  
445 新たな更新データ  
500 妥当性判定用データ  
510、520、530 各回目の更新に伴う冗長データ  
600 ファイル管理エントリ  
601 ファイル識別子  
620 レプリカ管理エントリ  
621 レプリカ識別子  
623 レプリカの存在アドレス  
624 レプリカの現バージョン  
625 レプリカの更新回数  
626 レプリカ更新データの妥当性判定用データ  
700 ファイルサーバにおけるレプリカ最新化処理  
750 レプリカサーバにおけるレプリカ最新化処理

15

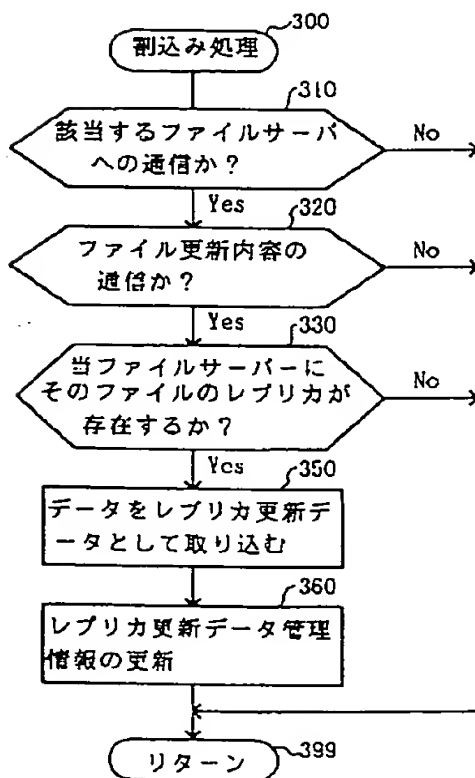
720 最新化要求に伴う通信データ

711、731 ファイルサーバからレプリカサーバへの通信

【図1】



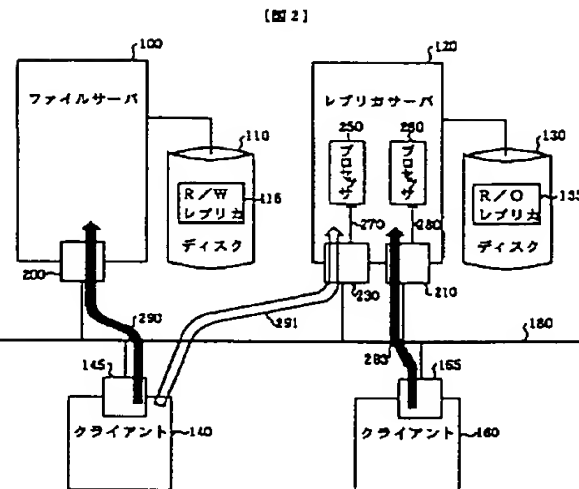
【図3】



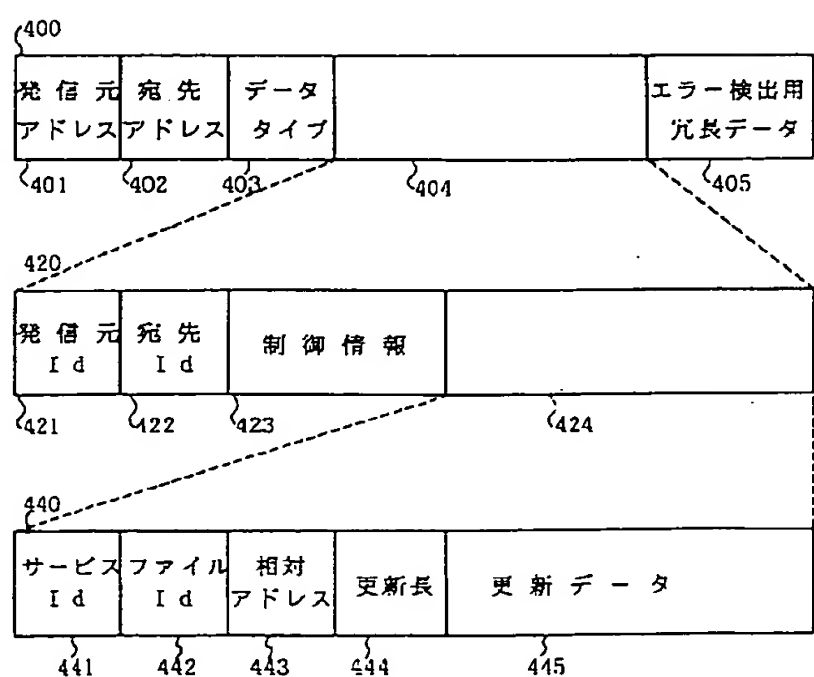
16

771、772 レプリカサーバからファイルサーバへの通信

【図2】

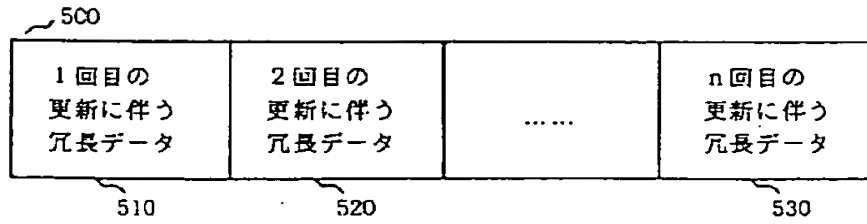


【図4】



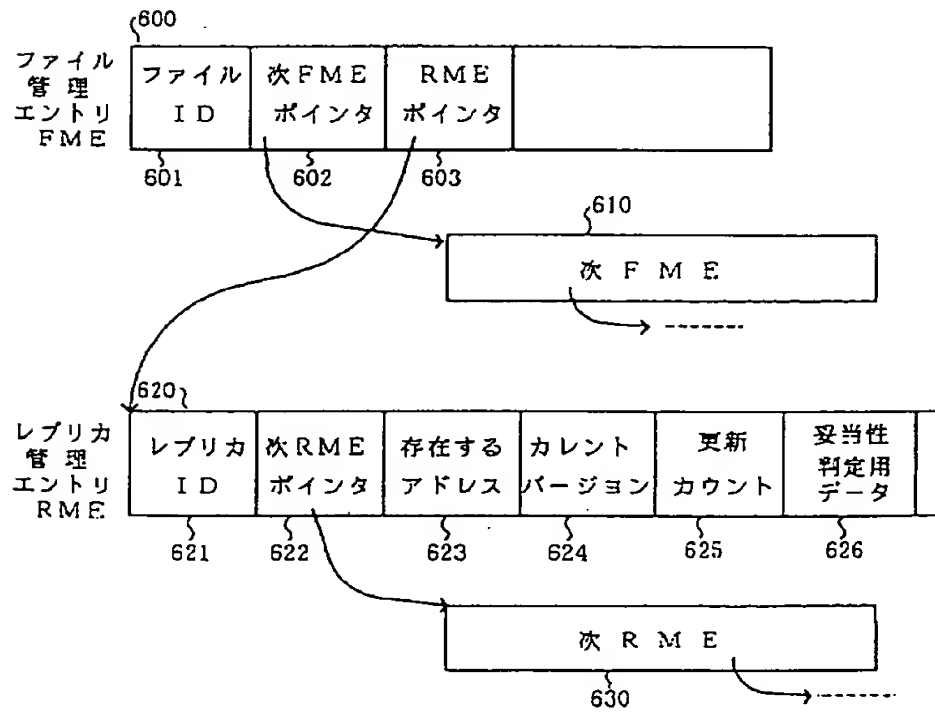
【図 5】

【図 5】



【図 6】

【図 6】



【図7】

【図7】

ファイルサーバにおける処理

レプリカサーバにおける処理

